DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

COMPARISON OF ROBUST REGRESSION METHODS

by
Gözde NAVRUZ

July, 2014
İZMİR

COMPARISON OF ROBUST REGRESSION METHODS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science in
Statistics, Statistics Program

by
Gözde NAVRUZ

July, 2014
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"COMPARISON OF ROBUST REGRESSION METHODS"** completed by **GÖZDE NAVRUZ** under supervision of **ASSOC. PROF. DR. A. FIRAT ÖZDEMİR** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....

Assoc. Prof. Dr. A. Fırat ÖZDEMİR

Supervisor

Prof. Dr. C. Cengiz Celikoğlu

(Jury Member)

Assoc. Prof. Dr. İlhan Karabekiroğlu

(Jury Member)

Prof. Dr. Ayşe OKUR

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to express my special appreciation and sincere thanks to my advisor Assoc. Prof. Dr. A. Fırat ÖZDEMİR, for his continuous support, patience and encouragement throughout my thesis. His guidance and help played a very important role in not only this study but also all my research.

I would like to acknowledge my dissertation committee members, Prof. Dr. C. Cengiz ÇELİKOĞLU and Assoc. Prof. Dr. İlhan KARAKILIÇ, for their constructive comments and suggestions.

I would like to express my heartfelt thanks to my parents Rakibe NAVRUZ and Coşkun NAVRUZ for their constant love plus moral and material support. They encouraged me to strive towards my goal and supported me in everything I decide to do. I also would like to thank my sister Ece NAVRUZ for her patience and invaluable love during this thesis. I would never succeed without their love and support.

Gözde NAVRUZ

COMPARISON OF ROBUST REGRESSION METHODS

ABSTRACT

Regression analysis is one of the most commonly used methods for examining data in many working areas. The ordinary least squares which is the conventional method used for estimating the parameters in a regression model can be misleading when there are outliers, heteroscedasticity and non-normality. In this study, after briefly explaining the problems with ordinary least squares general information about robustness idea is outlined and six robust regression techniques, Theil-Sen estimator, least median of squares, least trimmed of squares, least absolute value, least trimmed absolute value and M-regression that are not affected by these common problems are investigated.

The robust regression techniques are compared with a simulation study by using observations generated from both theoretical distributions and real data sets. The comparison is done in terms of actual significance levels and relative efficiency over ordinary least squares. Results and concluding remarks are discussed, some recommendations are given.

Keywords: Robust regression, M-regression, Theil-Sen, heteroscedasticity, non-normality

DAYANIKLI REGRESYON METOTLARININ KARŞILAŞTIRILMASI

ÖZ

Regresyon analizi, pek çok çalışma alanında verilerin incelenmesi için kullanılan en yaygın metotlardan biridir. Bir regresyon modelindeki parametreleri tahmin etmek için kullanılan geleneksel metot olan sıradan en küçük kareler yöntemi; aykırı değerler, heterojen varyanslılık ve normal olmayan dağılım durumlarında yanıltıcı olabilir. Bu çalışmada, sıradan en küçük kareler ile ilgili problemler özet olarak açıklandıktan sonra, dayanıklılık ile ilgili genel bilgi ana hatlarıyla belirtilmiş ve bu yaygın problemlerden etkilenmeyen altı tane dayanıklı regresyon tekniği, Theil-Sen kestiricisi, en küçük medyan kareler, en küçük budanmış kareler, en küçük mutlak değer, en küçük budanmış mutlak değer ve M-regresyonu incelenmiştir.

Dayanıklı regresyon teknikleri hem teorik dağılımlardan hem de gerçek veri setlerinden elde edilen gözlemlerin kullanıldığı bir simülasyon çalışması ile karşılaştırılmıştır. Karşılaştırma, gerçekleşen anlam düzeyleri ve sıradan en küçük kareler üzerindeki göreli etkinlik ölçütlerine göre yapılmıştır. Sonuçlar ve sonuç yorumları tartışılmış, bazı öneriler verilmiştir.

Anahtar kelimeler: Dayanıklı regresyon, M-regresyonu, Theil-Sen, heterojen varyanslılık, normal olmayan dağılım

CONTENTS **Page**

THESIS EXAMINATION RESULT FORM ii
ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZ v
LIST OF FIGURES viii
LIST OF TABLES ix

CHAPTER ONE - INTRODUCTION 1

CHAPTER TWO - ROBUSTNESS AND ROBUST REGRESSION TECHNIQUES 9

2.1 Robustness 9
2.1.1 Qualitative Robustness 9
2.1.2 Infinitesimal Robustness 10
2.1.1 Quantitative Robustness 10
2.2 Theil-Sen Estimator 11
2.3 Least Median of Squares 14
2.4 Least Trimmed Squares 15
2.5 Least Absolute Value 16
2.6 Least Trimmed Absolute Value 17
2.7 M-Regression 18

CHAPTER THREE - DESIGN OF SIMULATION STUDY AND RESULTS . 21

3.1 Design of Simulation 21
3.2 Results 26
3.2.1 Relative Efficiency Results 26
3.2.2 Actual Significance Levels Results 29
3.3 Real Data Example 34

CHAPTER FOUR - CONCLUSION 39

REFERENCES 42

LIST OF FIGURES

Page

Figure 1.1 Outliers in y-axis 4
Figure 1.2 Outliers in x-axis 4
Figure 1.3 Outliers in both x and y-axis 5
Figure 3.1 The shape of the g-and-h distribution when g=0 and h=0 22
Figure 3.2 The shape of the g-and-h distribution when g=0 and h=0.5 23
Figure 3.3 The shape of the g-and-h distribution when g=0.5 and h=0 23
Figure 3.4 The shape of the g-and-h distribution when g=0.5 and h=0.5 24
Figure 3.5 Actual significance levels of the three best robust regression estimators when x has normal distribution 30
Figure 3.6 Actual significance levels of the three best robust regression estimators when x has symmetric heavy tailed distribution 31
Figure 3.7 Actual significance levels of the three best robust regression estimators when x has asymmetric light tailed distribution 32
Figure 3.8 Actual significance levels of the three best robust regression estimators when x has asymmetric heavy tailed distribution 33
Figure 3.9 Histogram of total AA degrees 35
Figure 3.10 Histogram of total MA degrees 36
Figure 3.11 Histogram of total BA degrees 36
Figure 3.12 Histogram of GRE total 37

LIST OF TABLES

Page

Table 3.1 The skewness and kurtosis values of g-and-h distributions 22
Table 3.2 Estimated ratios of standard errors(R), x has normal distribution 26
Table 3.3 Estimated ratios of standard errors (R), x has symmetric heavy tailed distribution..... 27
Table 3.4 Estimated ratios of standard errors (R), x has asymmetric light tailed distribution..... 27
Table 3.5 Estimated ratios of standard errors (R), x has asymmetric heavy tailed distribution..... 28
Table 3.6 Actual significance levels, x has normal distribution 30
Table 3.7 Actual significance levels, x has symmetric heavy tailed distribution 31
Table 3.8 Actual significance levels, x has asymmetric light tailed distribution..... 32
Table 3.9 Actual significance levels, x has asymmetric heavy tailed distribution ... 33
Table 3.10 The descriptive statistics of real data sets 35
Table 3.11 Estimated ratios of standard errors (R) with real data 37
Table 3.12 Actual significance levels with real data..... 38

CHAPTER ONE
INTRODUCTION

Regression analysis is one of the most commonly used techniques for analyzing data. It expresses the relationship between a response variable which is referred to as Y and a set of predictor variables which are referred to as X. Note that when there is a unique predictor variable, the technique will be known as simple regression and if there are more than one predictor the technique will be multiple regression. Furthermore, when the predictions of Y values are plotted as a function of X values, if they form a straight line, the technique is called linear regression.

Regression analysis is commonly used in the areas like biological sciences, behavioral and social sciences, finance, economics and environmental sciences. Ordinary least squares (OLS) method is the general choice in a wide range of working areas. OLS estimator was discovered by Carl Friedrich Gauss in Germany about 1795 and by Adrien Marie Legendre in France about 1805, separately. First applications of the technique were done with astronomic and geodetic data. In 1805, in the appendix of a book by Legendre which is about determining the orbits of comets, OLS was published for the first time (Birkes & Dodge, 1993).

Even in those years, without computers, it was easy to compute OLS estimator and this makes it very popular among a lot of regression techniques.

Suppose $(y_i, x_{i1}, ..., x_{ip})$, i=1,2,...,n are vectors of p+1 variate distribution, where $(x_{i1}, ..., x_{ip})$ is the vector of predictors and suppose they are randomly sampled. The regression model can be formulated as

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \qquad (1.1)$$

where $\beta_0, ..., \beta_p$ are unknown parameters and ε_i are independent random variables with following assumptions:

- $E(\varepsilon_i) = 0$, the error term has zero mean.

- $\text{var}(\varepsilon_i) = \sigma^2$, the variance of the error term is constant.
- ε_i is independent of x_{ik} for each k, k=1,...,p.
- The errors are normally distributed.
- The errors are uncorrelated.
- The relationship between response variable and predictor variables is linear.

Note that, if $\text{var}(\varepsilon_i) = \sigma^2$,the model is called homoscedastic, that is the variance is constant. On the other side, if $\text{var}(\varepsilon_i) = \sigma_i^2$ and $\sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$, the model is called heteroscedastic.

Let

$$x_i = (x_{i1}, ..., x_{ip}) \tag{1.2}$$

and

$$\beta = (\beta_1, ..., \beta_p)' \tag{1.3}$$

$$= \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \tag{1.4}$$

Then we get

$$x_i\beta = \beta_1 x_{i1} + ... + \beta_p x_{ip} \tag{1.5}$$

and the equation (1.1) will be written as below by matrix notation:

$$y_i = \beta_0 + x_i\beta + \varepsilon_i \tag{1.6}$$

(Wilcox, 2012).

Let b_j be the estimate of β_j , j=1,...,p and let

$$\hat{y}_i = b_0 + b_1 x_{i1} + ... + b_p x_{ip} \tag{1.7}$$

where the estimates b_j values are known as regression coefficients, and $\hat{y}_i$ is called the predicted or estimated value of y_i . Then the residual r_i is defined as

$$r_i = y_i - \hat{y}_i \tag{1.8}$$

Note that the residual r_i is generally denoted with $\hat{\varepsilon}_i$ in the literature, and in this thesis the notation is chosen as r_i. That is, the residual is the difference between the observed value and estimated value (Rousseeuw & Leroy,1987).

The ordinary least squares(OLS) estimator aims to minimize

$$\sum r_i^2 \tag{1.9}$$

As an example consider simple regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \tag{1.10}$$

The OLS estimator of the slope β_1 is

$$\hat{\beta}_1 = \frac{\sum (x_{i1} - \bar{x}_1)(y_i - \bar{y})}{\sum (x_{i1} - \bar{x}_1)^2} \tag{1.11}$$

$$\bar{x}_1 = \frac{\sum x_{i1}}{n} \tag{1.12}$$

$$\bar{y} = \frac{\sum y_i}{n} \tag{1.13}$$

and corresponding OLS estimate of the intercept β_0 is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 \tag{1.14}$$

However, OLS gives correct results if the assumptions of the model are hold. Otherwise, OLS can be misleading. Unfortunately, the real data generally do not satisfy the basic assumptions such as normality, homoscedasticity etc. Also, if there are outliers in data, the assumption of normality is violated.

Outliers are data points that are far away from the bulk of data. In other words, outliers are the unusual observations that have extreme values on one or more variables. The reasons of outliers can be data entry errors, scoring errors or the sample data can be biased. When their location is considered, outliers can be in x-axis, y-axis or both of them.

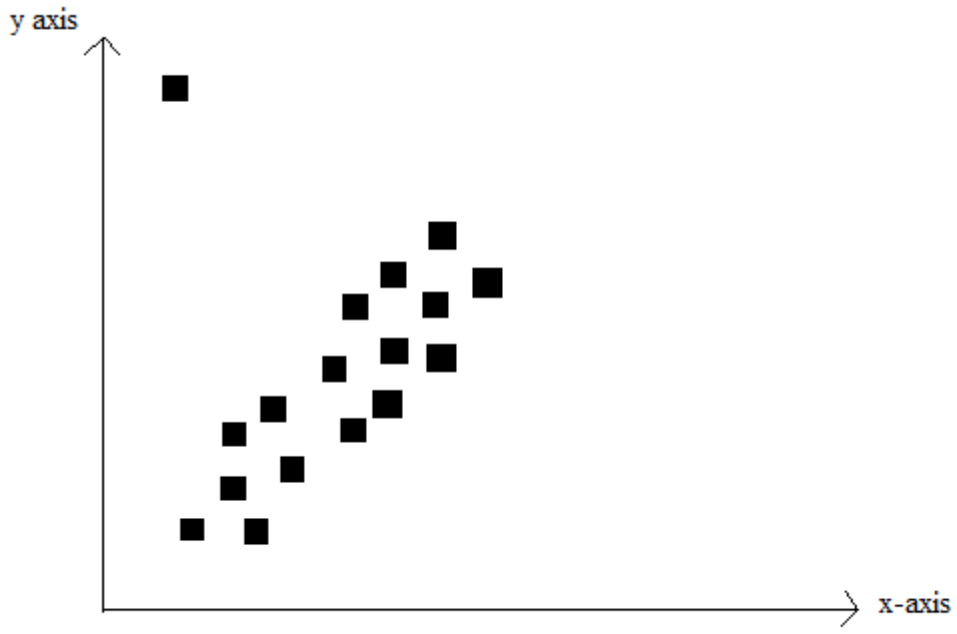


Figure 1.1 Outliers in y-axis

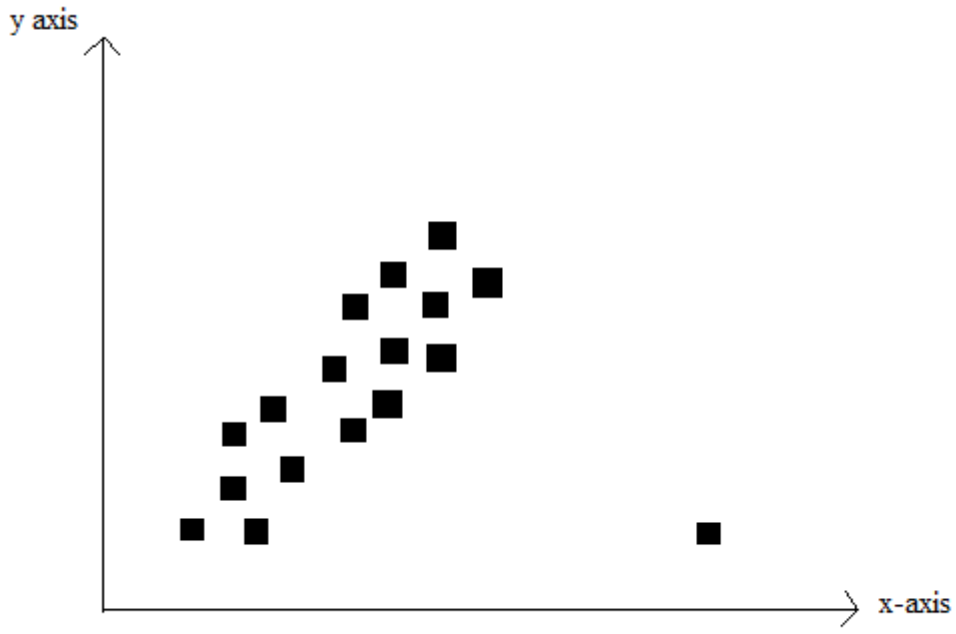


Figure 1.2 Outliers in x-axis

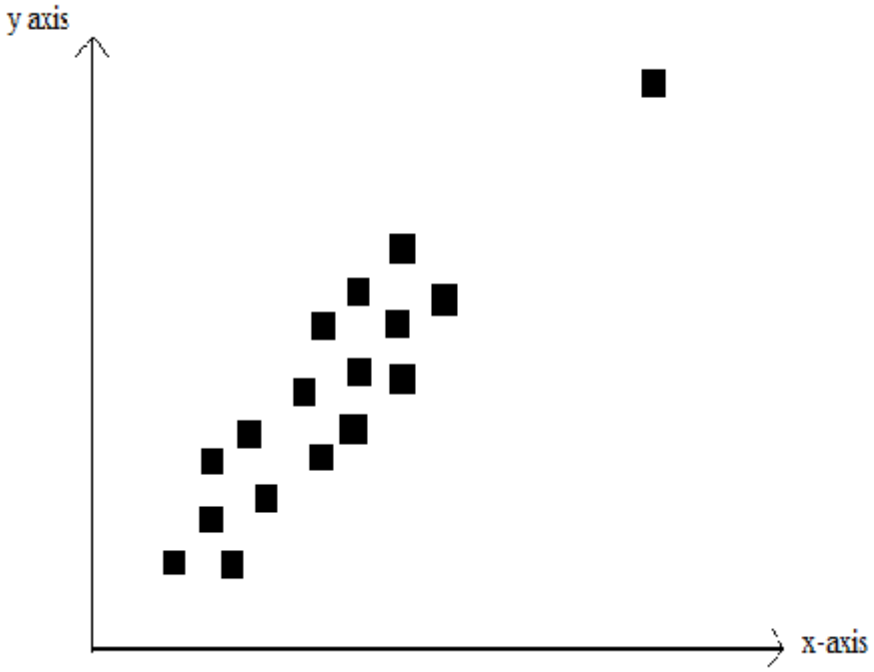


Figure 1.3 Outliers in both x and y-axis

Outlying x values are called leverage point, because of its resemblance to the fact of leverage in mechanics and their effect on regression estimators is large (Rousseeuw & Leroy,1987). Furthermore there are two types of leverage point, good or bad. Being good or bad is determined with respect to their consistency with true regression line. When the leverage point stays close to the regression line obtained by the bulk of the data, it is considered as a good leverage point. A point which have a relatively large residual is called regression outlier. A good leverage point is not a regression outlier, however a bad leverage point is not only a leverage point but also a regression outlier. Also, a good leverage point can make the standard error of estimator is smaller (Rousseeuw & Leroy, 1987; Wilcox, 2012).

If normality or homoscedasticity assumptions are not hold, some problems arise with conventional technique OLS. First of all, the ordinary least squares estimator can have a relatively high standard error, that means OLS estimator can be highly inefficient. This problem may occur when the error term has a normal distribution

with heteroscedasticity. On the other hand it may occur if the error term is homoscedastic, but it has a heavy tailed distribution. Furthermore, low efficiency leads low power (Wilcox, 1996a, 2012; Wilcox & Keselman, 2012).

When the assumptions are violated, another problem is about confidence intervals. The probability coverage of confidence intervals can be highly different from the nominal level. When there is heteroscedasticity, the actual significance level may exceed 0.5, where the nominal level is 0.05 (Wilcox 1996a, 1996b). An important effect of this problem is that, when ordinary least squares technique says the result is highly non-significant, robust regression techniques can detect a strong association. That is, the conventional model cannot reflect the form of the association (Wilcox & Keselman, 2012).

Furthermore, ordinary least squares estimator has a finite sample breakdown point of 1/n, meaning that a single observation can make its value considerably larger or smaller. It is important to note that, not only outlying y values, but also outlying x values can cause substantially smaller or larger estimates (Wilcox, 2012).

To cope with given problems, an usual strategy is to check for outliers and after removing them, use the conventional method with the remainder data. However, in this case, another problem arises; the estimated standard error may not converge to the correct value as the sample size increases. Because, the standard error of the ordinary least squares estimator is obtained with the assumption of that the observations are independent. But, removing outliers make the remainder observations dependent. A bootstrap estimate of the standard error can be thought as a solution, but there is not enough knowledge about its effectiveness. Besides, easily removing outliers does not solve low efficiency problem due to heteroscedasticity (Wilcox, 2012).

Another strategy is to test the assumptions of homoscedasticity and normality. Also, there is a wide range of such tests. But, most of these tests are not good at controlling probability of a Type I error. On the other hand, the tests that can control

the probability of a Type I error have another incompetence; their power is not enough to determine the cases where heteroscedasticity cause poor probability coverage (Wilcox & Keselman, 2012).

Because of the given problems, more suitable approach is to use a technique that performs better when there is heteroscedasticity, and that can compete with the conventional technique when the model is homoscedastic (Wilcox, 2012).

Since the aim is to find the associations between random variables and make inferences about regression parameters, instead of ordinary least squares, more modern techniques that allow violation of normality and homoscedasticity assumptions must be used. These modern techniques are known as robust regression techniques. Also, robust regression techniques overcome outliers effectively.

Basically, robust methods are insensitive to small alteration from the assumptions (Huber, 1981). The fundamental information and development of robustness theory are explained in the books by Huber and by Hampel et al (Hampel, Ronchetti, Rousseeuw & Stahel, 1986; Huber, 1981).

Additionally, more information about robust regression can be found in the books by Rousseuw & Leroy and by Birkes & Dodge (Birkes & Dodge, 1993; Rousseuw & Leroy, 1987). Furthermore, Wilcox briefly explained the idea of robustness and robust regression techniques (Wilcox, 2012).

Although there are many robust regression methods, none of them performs better than the others in all conditions. So, it is important to know their procedures and properties. The aim of this study is to define and compare some of the robust regression techniques that are commonly accepted in the literature. These techniques are; Theil-Sen estimator, least median of squares, least trimmed squares, least absolute value, least trimmed absolute value and M-regression. Also, the reason of using these techniques is explained in this thesis. The comparison of robust regression techniques is done in terms of their relative efficiency over ordinary least

squares regression and actual significance levels. Results are discussed and some recommendations are given.

This thesis is organized in four chapters. In the second chapter; general information about robustness theory is given and the robust regression techniques Theil-Sen estimator, least median of squares, least trimmed squares, least absolute value, least trimmed absolute value and M-regression estimator are defined. The design of the simulation study is stated in the chapter three. To compare the robust regression techniques in terms of their relative efficiency over ordinary least squares regression and in terms of their actual significance levels, a simulation study is performed. Furthermore, the results of the simulation study is given in chapter three. Finally, conclusions and some recommendations for future studies are given in the chapter four.

CHAPTER TWO
ROBUSTNESS AND ROBUST REGRESSION TECHNIQUES

2.1 Robustness

Robust statistics is related with mostly seen assumptions in statistics. The assumptions can be about randomness, independence, linearity and as well normality or homoscedasticity. Unfortunately, the assumptions do not always hold in real world (Hampel et al, 1986; Huber, 1981).

The robustness problem is based to early years of statistics and also some important statisticians were conscious of this. However, the number of people who consider this problem has started to increase in recent decades. Hence it became essential to formalize the definition of robustness (Hampel et al, 1986; Huber, 1981).

Additionally the term robustness was used for the first time by Box in 1953 (Box, 1953; David, 1995).

In short, robustness means insensitivity to slight departures from assumptions or their violations. There are three basic types of robustness; qualitative robustness, quantitative robustness and infinitesimal robustness.

2. 1. 1 Qualitative Robustness

Let θ be any measure about some feature of a distribution, and let F and G be any two distributions. The value of θ, when sampling from F is shown as $\theta(F)$. A main necessity for robustness is continuity. That is, if F is close to G, $\theta(F)$ must be close to $\theta(G)$ (Wilcox, 1998a).

To define qualitative robustness in a formal way, let $G_1, G_2, \ldots$ be any sequence of distributions which converges to the distribution F (assumed model). If $\theta(G_1), \theta(G_2), \ldots$ converge to $\theta(F)$, then θ is qualitative robust (Wilcox, 1998a).

As an example, the arithmetic mean is not qualitative robust. On the other hand, the median and α-trimmed mean $(0 < \alpha < 0.5)$ are qualitative robust (Hampel et al, 1986).

2. 1. 2 Infinitesimal Robustness

In robust statistics literature, parameters are viewed as functionals so that the notion of continuity and boundedness can be extended to them. A functional is a rule or formula which maps distributions to real numbers. The other type of robustness is infinitesimal robustness which is an issue concerning influence function. The influence function of a functional T at assumed distribution (model) F is the derivative of the functional T(F) and it measures the influence of small changes in assumed model F (Hampel, 1974).

The influence function expresses the influence of adding another one observation to a very large sample. In another words, it examines the influence of a new observation which is added, over the estimated value of any measure or test statistic. If the influence function is not bounded, an outlier can make the estimator arbitrarily larger or smaller (Huber, 1981). Parameters having a bounded influence function are said to have infinitesimal robustness property against outliers.

2. 1. 3 Quantitative Robustness

The quantitative robustness is known with the term breakdown point which is the limiting value of finite sample breakdown point. The aim is to explain how a slight variation in F effects T(F), quantitatively (Wilcox, 2012).

The finite sample breakdown point is the proportion of observations, that when changed, can make the value of estimator larger or smaller. In other words, finite sample breakdown point says, an estimator can cope with how many outliers. The necessity is that, the breakdown point of a parameter is greater than zero to say a parameter is quantitatively robust (Wilcox, 1998a).

The smallest possible value of breakdown point is 0 and the largest possible value is 0.5. For example, the arithmetic mean has finite sample breakdown point of 1/n, which means only one outlier can devastate its value. Also, taking the limit its breakdown point is found as 0 (Wilcox, 1998a).

As an another example, the sample median has a breakdown point of 0.5. If an estimator has breakdown point of 0.5, it means fifty percent of the data can contain outliers and the value of the estimator does not affected.

2.2 Theil-Sen Estimator

The Theil-Sen estimator first proposed by Theil (1950) and later improved by Sen(1968). The estimator which is found by Theil, is not defined if there are ties among x values; Sen extended this estimator to handle ties among x values. Firstly, when the idea is focused on single predictor (p=1), Theil-Sen estimator aims to find a value for the slope that makes Kendall's correlation tau between $y_i - bx_i$ and x_i approximately equal to zero. For any $i < i'$, for which $x_i \neq x_i'$, let

$$s_{ii'} = \frac{y_i - y_i'}{x_i - x_i'} \tag{2.1}$$

That is, all possible pairwise slopes are evaluated. The Theil-Sen estimate of the slope b_{1ts} is the median of all slopes represented by $s_{ii'}$. The corresponding intercept estimate is found by $M_y - bM_x$,where M_y and M_x are the usual sample medians of the y and x values, respectively. Also, Theil-Sen estimator is an unbiased estimator for the slope (Sen, 1968).

There are a number of ways of extending the Theil-Sen estimator to two or more predictors. The first one of such techniques, is known as TS method. TS method applies back fitting, Gauss Seidel method. The steps of TS methods are performed as follows:

(1) Set k=0 and select an initial estimate for β_j, say $b_j^{(0)}$, j=0,...,p. Now, the initial estimate is the Theil-Sen estimate of the slope evaluated on the basis of the only jth predictor . Namely, the other predictors are disregarded to get an initial estimate of β_j. The initial estimate of β_0 is the median of

$$y_i - \sum_{j=1}^{p} b_j^{(0)} x_{ij} \quad , i=1,...,n. \tag{2. 2}$$

(2) Iterate, where kth estimate $b_j^{(k)}$,is the Theil-Sen estimate of the slope on the basis of the regression estimate of x_{ij} with

$$r_i = y_i - b_0^{(k-1)} - \sum_{l=1,l\neq j}^{p} b_l^{(k-1)} x_{il} \tag{2.3}$$

The new corresponding estimate of the intercept is the median of

$$y_i - \sum_{j=1}^{p} b_j^{(k)} x_{ij} \quad , i=1,...,n. \tag{2.4}$$

(3) Repeat step two up t convergence (Wilcox, 2012).

TS technique is regression equivariant and scale equivariant, but it is not affine equivariant (Wilcox, 2012).

A regression estimator T is affine equivariant if for any nonsingular matrix A,

$$T(x_i A, y_i; i=1,...,n) = A^{-1} T(x_i, y_i; i=1,...,n) \tag{2.5}$$

A regression estimator T is regression equivariant if for any vector v,

$$T(x_i, y_i + x_i v; i=1,...,n) = T(x_i, y_i; i=1,...,n) + v \tag{2.6}$$

A regression estimator T is scale equivariant if,

$$T(x_i, c\,y_i; i = 1, ..., n) = cT(x_i, y_i; i = 1, ..., n) \quad (2.7)$$

The second method of extending the Theil-Sen estimator to two or more predictor is called TSG and is based on the idea known as elemental subsets (Hawkins & Olive, 2002; Rousseeuw & Leroy, 1987). An elemental subset contains minimum number of cases needed to estimate the unknown parameters of a regression model. The elemental subset idea was recommended for the first time by Oja and Niimimaa in an unpublished work and according to their extension all $N = n!/[(p+1)!(n-p-1)!]$ elemental subsets are used with p regressors. For each elemental subset, find the estimates of the parameters by using OLS, then use the median of N present estimates. But, there is a problem that, N increases fast with the increase of n and p. So, another idea is to use $(n^2 - n)/2$ elemental subsets instead(Wilcox, 2012). Furthermore, in terms of efficiency using $(n^2 - n)/2$ elemental subsets gives better results than using all N elemental subsets (Wilcox, 1998b). Note that, TSG technique is not affine equivariant and as p increases, the breakdown point decreases with this method.

A third approach is to make $\sum|\hat{\tau}_j|$ is approximately equal to zero, where $\hat{\tau}_j$ is the Kendall's tau between x_j and $y - b_1x_1 - ... - b_px_p$. Here, another coefficient of correlation can also be used.

The choice of the method can make a practical difference, for example TS works faster than TSG. But none of these techniques is preferable over the others (Wilcox, 2012).

Theil-Sen estimator has a bounded influence function. It has a breakdown point 0.293 which is high enough to protect the estimates from outliers (Dietz, 1989). Under both normality and homoscedasticity, the efficiency of Theil-Sen estimator can compete well with the OLS. Furthermore, if there is normality under heteroscedasticity, the efficiency of Theil-Sen estimator is much better (Wilcox, 2010).

2.3 Least Median of Squares

The least median of squares (LMS) regression estimator was proposed by Rousseeuw(1984) to obtain an estimator that has a high breakdown point. Instead of minimizing the sum of the squared residuals, the median of the squared residuals is used. The regression estimates are the values that minimize the median of the squared residuals, $\mathrm{MED}(r_1^2,...,r_n^2)$ (Rousseeuw, 1984).

The idea is based on sampling a large number of elemental subsets and estimating the slope and regression coefficients for each of them. Then by using all estimates the median of the squared residuals is evaluated and the one that minimizes the median squared residual is chosen (Rousseeuw & Leroy, 1987).

When $p>1$, the finite sample breakdown point of the LMS technique is $([n/2]-p+2)/n$, note that [x] means the greatest integer less than or equal to x. The breakdown point is based on n, and if a single value is considered, the limit when $n\to\infty$ is taken and the breakdown point is found as 0.5 (Rousseeuw, 1984). LMS estimator is the first equivariant estimator that has a breakdown point 0.5, the maximum possible value. Besides, it has the largest breakdown point among all regression equivariant estimators. As well as regression equivariance, LMS is scale equivariant and affine equivariant (Rousseeuw & Leroy, 1987).

However, it has a very low efficiency, it has asymptotic efficiency of %0 (Stromberg, Hössjer & Hawkins, 2000). Because the rate of convergence of LMS is $n^{-1/3}$, rather than the usual rate of $n^{-1/2}$. Still, the low efficiency is not a big problem since LMS is generally thought as a data analytic tool and in such cases efficiency is not the most important criteria (Rousseeuw, 1984).

LMS technique minimizes the median of the squared residuals and ignores the remainder of the data points, so it is not recommended as a single procedure. Furthermore, LMS is not computationally practical, because it evaluates all possible subsets to estimate the slope and or regression coefficients. Another negative feature

is its low efficiency. Despite these negative properties, the LMS estimator can be useful in detection of outliers, it can be used as a diagnostic tool for data or it can be a robust initial estimate for iterative process (Birkes & Dodge, 1993; Wilcox, 1996c).

2.4 Least Trimmed Squares

To improve the efficiency of LMS, Rousseeuw (1984) defined the least trimmed squares (LTS) estimator. LTS estimator aims to minimize

$$\sum_{i=1}^{h} r_{(i)}^2 \tag{2.8}$$

where $r_{(1)}^2 \leq \ldots \leq r_{(h)}^2$ are the squared residuals written in increasing order. An important difficulty of LTS estimator is that, its evaluation needs sorting of the squared residuals and this cause extremely large computation times (Rousseeuw, 1984).

The main difference between OLS and LTS estimators is minimizing the sum of first h squared residuals instead of all residuals. Thus, deciding the number h is important. With the choice

$$h = [n/2] + 1 \tag{2.9}$$

LTS estimator has a finite sample breakdown point of $([n/2] - p + 2)/n$, the same value which the LMS estimator has. With regard to breakdown point, LTS is thought as a high breakdown method with 0.5 and hence it can handle outliers. If,

$$h = [n/2] + [(p+1)/2] \tag{2.10}$$

LTS will be regression equivariant (Wilcox, 2012). In general, h may depend on the trimming proportion α, where $0 < \alpha < 0.5$, that is

$$h = n(1-\alpha) + 1 \tag{2.11}$$

$$h = [n(1-\alpha)] + [\alpha(p+1)] \tag{2.12}$$

are the possible values which contains trimming proportion in the formula. In such cases, the breakdown point of the LTS estimator is equal to the trimming proportion α. When α is close to 0, again the OLS estimator is obtained (Rousseeuw & Leroy, 1987).

LTS has the usual rate of convergence $n^{-1/2}$, so its efficiency is better than LMS. However, still LTS has a relatively low asymptotic efficiency (Croux, Rousseeuw & Hössjer, 1994). Its asymptotic efficiency is only 0.08 (Stromberg et al, 2000). Additionally, the LTS estimator not only regression equivariant, but also scale equivariant and affine equivariant (Rousseeuw & Leroy, 1987).

LTS method is a reliable data analysis tool because its robustness does not affected by the number of independent variables, so can be used in multiple regression case. However, it is not acceptable as a single procedure. It can be useful in the calculation of the other estimators or it can be used as initial estimates. Moreover, it is used effectively in the outlier detection.

2.5 Least Absolute Value

Least absolute value (LAV) approach was developed in 1757, almost 50 years before the method of OLS. It was first defined by Roger Joseph Boscovich. His purpose was to improve the conflicting results with the aim of estimating the shape of earth. After Boscovich, Pierre Simon Laplace improved and used it. This method was rarely used and by the occurrence of OLS, this method took a secondary place (Birkes & Dodge, 1993).

LAV estimator aims to find the estimates of slope and intercept that minimize the sum of the absolute value of residuals

$$\sum_{i=1}^{n} |r_i| \tag{2.13}$$

LAV regression is also known as L_1 regression, because L_1 norm is used to define it; where

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \tag{2.14}$$

is the L_1 norm of the vector $x = (x_1, x_2, ..., x_n)$. Furthermore, OLS regression is known as L_2 regression, because it minimizes the L_2 norm, where

$$\|x\|_2 = (\sum_{i=1}^{n} x_i^2)^{1/2} \quad (2.15)$$

is the L_2 norm of the vector $x = (x_1, x_2, ..., x_n)$ (Birkes & Dodge, 1993).

The LAV estimator is robust against the outliers among the Y values, but it is still highly sensitive to the outliers among X values. Although it reduces the effects of outliers, it has a finite sample breakdown point $1/n$ (Wilcox, 2012). It can have advantages in terms of relative efficiency over OLS. However, it is neither a bounded influence nor a high breakdown point estimator.

Although LAV estimator was found before OLS, it did not become popular; because instead of a formula, there are algorithms for calculating the LAV estimator and its algorithms needs a long computation time. Another problem arises about uniqueness, that means there may be exist more than one best line passing through a data point (Birkes & Dodge, 1993).

2.6 Least Trimmed Absolute Value

Least trimmed absolute (LTA) value estimator was developed in the early 1990s. LTA estimator aims to find the slope and intercept that minimizes the sum of the first h absolute residuals when the absolute residuals are in ascending order. That is, LTA estimator is found by minimizing

$$\sum_{i=1}^{h} |r|_{(i)} \quad (2.16)$$

where $|r|_{(i)}$ is the ith smallest absolute residual.

This estimator is a special case of LAV regression estimator with the difference of trimming. There is two important matter when finding the LTA estimator; deciding the value of the h and determining the elemental subsets that is used in minimizing L_1 norm. In general

$$h = [(n+p+1)/2] \quad (2.17)$$

is chosen and this value maximizes the breakdown point. As LMS and LTS estimators, LTA is a high breakdown method with 0.5 (Hawkins & Olive, 1999).

The calculation of LTA is based on elemental regression. That is;

- Create each elemental subset,
- For every elemental subset, fit the regression function and evaluate the residuals for each case,
- Decide on the smallest h absolute values of residuals and find their sum,
- LTA estimator is obtained by the elemental fit that makes the sum of the first h absolute residuals is smallest (Hawkins & Olive, 1999).

LTA estimator is a desirable alternative to LMS and LTS, especially for large data sets. Its efficiency is not much below than the efficiency of LTS for outlier-free normal data. Furthermore it is better than LTS for more peaked error distributions. Besides, computation of LTA is simpler than LMS and LTS, using for larger data set is easier (Hawkins & Olive, 1999).

2.7 M-Regression

Huber generated a group of estimators called M-estimators (Huber, 1981). Regression M-estimators corresponds to a generalization of the location M-estimators. M-estimators of regression are based on choosing a function ξ and then estimating β_j, (j=0,...,1) with b_j values that minimize

$$\sum \xi(r_i) \tag{2.18}$$

Generally ξ has a unique minimum at zero and it is some symmetric function, that is

$$\xi(r) = \xi(-r) \tag{2.19}$$

(Rousseuw & Leroy, 1987). As a special case, to get ordinary least squares this function is chosen as,

$$\xi(r) = r^2 \tag{2.20}$$

and to get least absolute value estimator this function is chosen as,

$$\xi(r) = |r| \tag{2.21}$$

(Birkes & Dodge, 1993).

Here, Huber's function is considered, where Huber's ψ is

$$\psi(x) = \max[-K, \min(K, x)] \tag{2.22}$$

or it can be written as

$$\psi(x) = \begin{cases} x & , |x| \le K \\ K\,\text{sign}(x) & , |x| > K \end{cases} \tag{2.23}$$

and ψ is the derivative of ξ , and corresponding ξ function is

$$\xi(x) = \begin{cases} \frac{1}{2}x^2 & , |x| \le K \\ |x|K - \frac{1}{2}K^2 & , |x| > K \end{cases} \tag{2.24}$$

The value of the constant K can be taken as

$$K = 2\sqrt{(p+1)/n} \tag{2.25}$$

(Wilcox, 2012).

To get scale equivariance, a measure of scale is needed in the sum that will be minimized, and with the measure of scale τ, M regression estimators now minimize

$$\sum \xi\left(\frac{r_i}{\tau}\right) \tag{2.26}$$

In addition, the measure of scale in the formula can be estimated by

$$\hat{\tau} = \frac{\text{median of the l largest } (n-p-1) \text{ of the } |r_i|}{0.6745} \tag{2.27}$$

(Wilcox, 2012). Now, an iterative estimation is used to obtain M- regression estimator, which is the solution of the following system of p+1 equation

$$\sum_{i=1}^{n} x_{ij} w_i r_i = 0 \ , \quad j=0,...p \tag{2.28}$$

where $x_{i0} = 1$ and

$$w_i = \begin{cases} \psi(r_i / \hat{\tau}) & \text{if} \quad y_i \neq \hat{y}_i \\ 1 & \text{if} \quad y_i = \hat{y}_i \end{cases} \tag{2.29}$$

(Wilcox, 2012). To start the algorithm, the initial estimators of β_j are found by ordinary least squares. However, by using more robust estimators such that LAV or LMS instead of OLS, the convergence of the algorithm may be simplified (Birkes & Dodge, 1993).

M- estimators are more efficient than LAV estimators. Furthermore, it is robust against outlying y values, but it is not robust against leverage points, and this is why it has still 1/n finite sample breakdown point (Rousseeuw & Leroy, 1993).

CHAPTER THREE
DESIGN OF SIMULATION STUDY AND RESULTS

3.1 Design of Simulation

In this chapter of the thesis, six robust regression techniques Theil-Sen estimator, least median of squares, least trimmed squares, least absolute value, least trimmed absolute value and M-regression and conventional technique ordinary least squares are compared with a simulation study. There is one predictor x, so only simple regression is considered. Furthermore the distributions of x and the error term ε were taken as standard normal distribution and g-and-h distribution.

g-and-h distribution allows to observe how a distribution differs from normal distribution with the parameters g, which is about skewness and with a parameter h, which is about kurtosis; and when g and h are both equal to zero, the g-and-h distribution is equivalent to standard normal distribution (Hoaglin, 1985).

Let Z be a random variable which is generated from standard normal distribution. When $g \neq 0$, with the transformation

$$X = \frac{(\exp(gZ) - 1)\exp(\frac{hZ^2}{2})}{g} \tag{3.1}$$

and when g=0, with the transformation

$$X = Z\exp(\frac{hZ^2}{2}) \tag{3.2}$$

the data is generated from g-and-h distribution (Hoaglin,1985).

In this thesis, in addition to standard normal distribution, three different g-and-h distributions are used:

- Symmetric heavy tailed (SH) g-and-h distribution with g=0 and h=0.5
- Asymmetric light tailed (AL) g-and-h distribution with g=0.5 and h=0

- Asymmetric heavy tailed (AH) g-and-h distribution with g=0.5 and h=0.5

The skewness and kurtosis values of these distributions are given in the Table 3.1 where they are estimated with $\hat{\kappa}_1$ and $\hat{\kappa}_2$ respectively. These estimated skewness and kurtosis values are based on 100000 observations. When g=0, it is known that skewness equals to zero, so in such cases skewness is not estimated (Wilcox, 2012).

For g and h values in Table 3.1, the shapes of the distributions are given in Figure 3.1, Figure 3.2, Figure 3.3 and Figure 3.4.

Table 3.1 The skewness and kurtosis values of g-and-h distributions

g	h	$\hat{\kappa}_1$	$\hat{\kappa}_2$
0	0	0	3
0	0.5	0	11986.2
0.5	0	1.81	9.7
0.5	0.5	120.1	18393.6

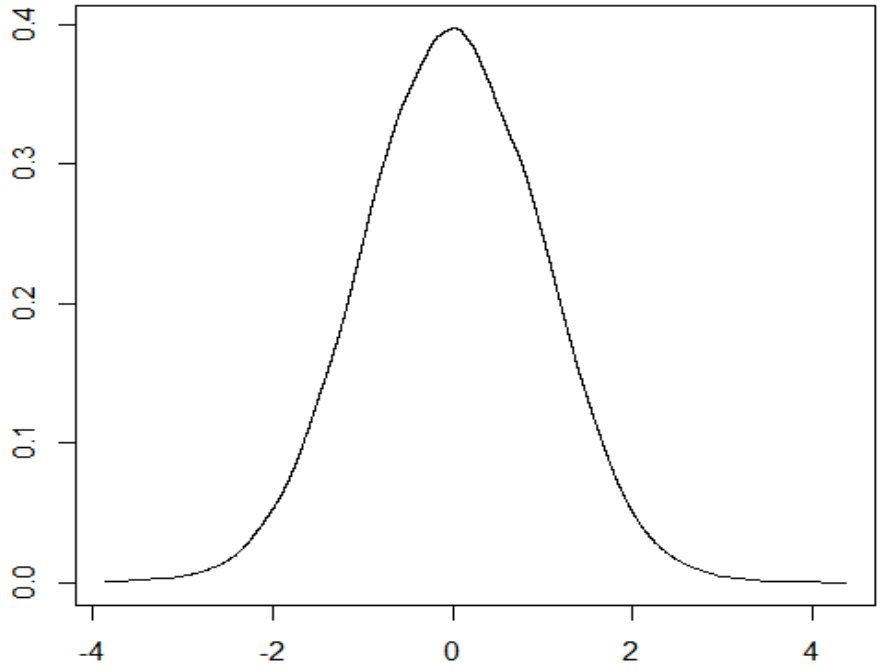


Figure 3.1 The shape of the g-and-h distribution when g=0 and h=0

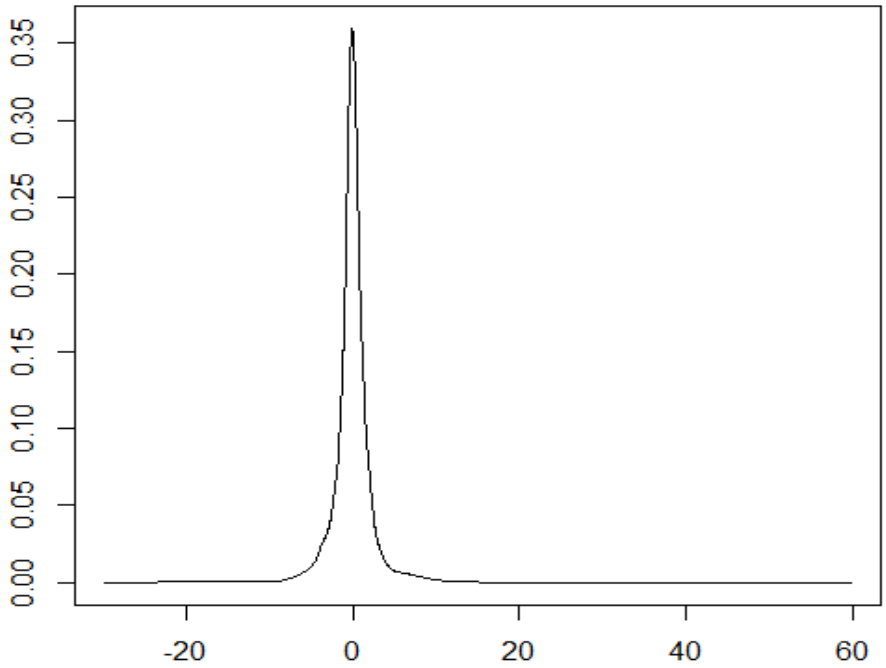


Figure 3.2 The shape of the g-and-h distribution when g=0 and h=0.5

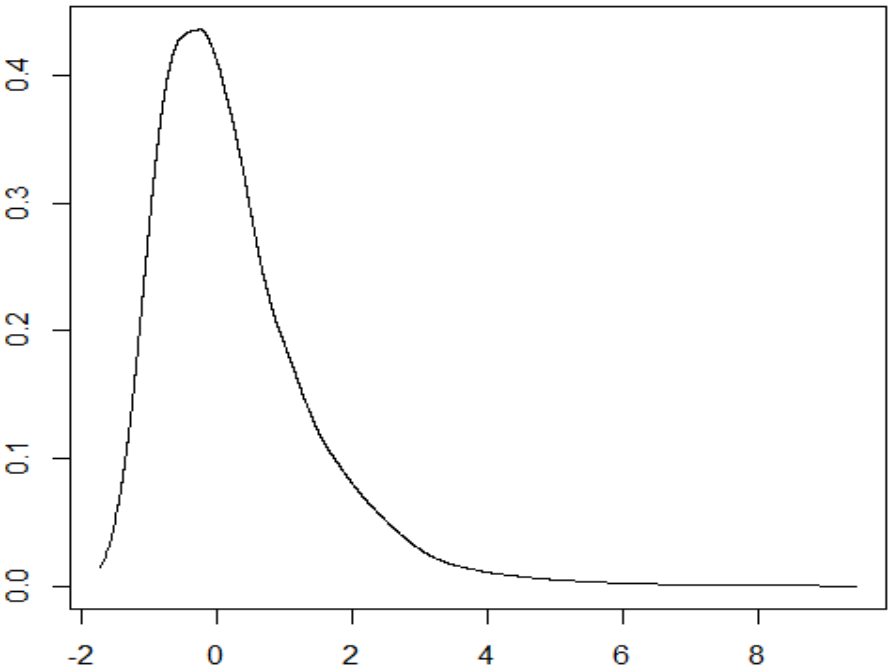


Figure 3.3 The shape of the g-and-h distribution when g=0.5 and h=0

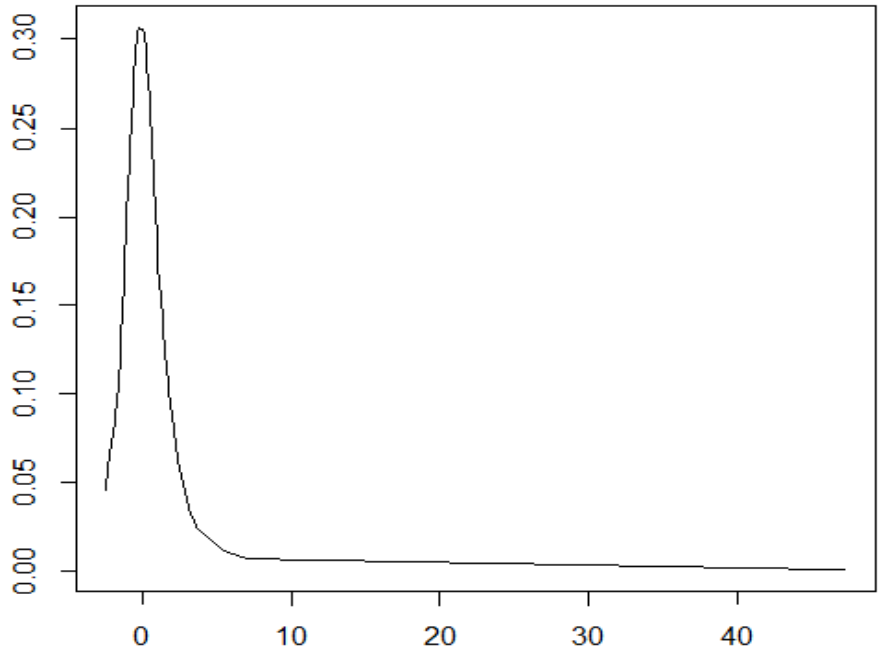


Figure 3.4 The shape of the g-and-h distribution when g=0.5 and h=0.5

The robust regression techniques are compared according to actual significance levels($\hat{\alpha}$) and their relative efficiency over ordinary least squares. Relative efficiency is measured by a the ratio R, where R is the standard error of ordinary least squares estimator is divided by standard error of a robust regression estimator. If this ratio is greater than one, it means ordinary least squares estimator has a larger standard error and so the robust estimator will be more appropriate (Wilcox, 2012).

$$R = \frac{s.e(\hat{\theta}_{OLS})}{s.e(\hat{\theta}_{robust})} \tag{3.3}$$

According to the model that observations are generated, a function $\lambda(x)$ reflects the heteroscedasticity. At first, to compare the robust regression estimators in terms of relative efficiency, the observations are generated with the model

$$y = x + \lambda(x)\varepsilon \tag{3.4}$$

And for this simulation there are three different variance patterns VP1, VP2 and VP3. VP1 corresponds to $\lambda(x)=1$, VP2 corresponds to $\lambda(x)=x^2$ and VP3 corresponds to $\lambda(x)=1/|x|$. VP1 equals to homoscedastic case. For VP2, the error term ε has more variance than extreme x values and on the other hand, for VP3 the error term ε has smaller variance than extreme x values (Wilcox, 2012)

Moreover, for comparing estimators in terms of actual significance levels($\hat{\alpha}$), again three variance patterns are used where VP1 corresponds to $\lambda(x)=1$ and VP2 corresponds to $\lambda(x)=x^2$, but now VP3 equals $\lambda(x)=1+2/(|x|+1)$. When as comparing with respective to relative efficiency, here VP2 means the error term ε has more variance than extreme x values, and as well VP3 means the error term ε has smaller variance than extreme x values. For this part of the simulation study the observations are generated with the model (Wilcox, 2012);

$$y=\lambda(x)\varepsilon \tag{3.5}$$

The sample size of n=20 are taken from each distribution and the nominal significance level is taken as $\alpha=0.05$. R programming language (R version 3.0.0) is used. Simulations are based on 5000 replications for actual significance level and based on 10000 replications for relative efficiency. In terms of relative efficiency, an estimator is considered as robust if the ratio R is greater than one. Additionally, an estimator is considered as robust if its actual significance level falls in the interval (0.025, 0.075). These bounds are formed according to Bradley's liberal criteria (Bradley, 1978). Also note that, the robust regression estimators Theil-Sen estimator, least median of squares, least trimmed squares, least absolute value, least trimmed absolute value and M-regression are denoted with TS, LMS, LTS, LAV, LTA and M respectively, as from this part. Furthermore ordinary least squares is represented with OLS.

3.2 Results

3.2.1 Relative Efficiency Results

In terms of relative efficiency the results are given in Table 3.2, Table 3.3, Table 3.4 and Table 3.5. In each case, the greatest three R ratio is marked and the robust regression estimators that corresponds these marked R ratios are given as preference in last columns of the tables. The tables are organized according to the distribution of x. For example, in Table 3.2, x has standard normal distribution and the results are given for each error distribution N, SH, AL, AH and each variance patterns VP1, VP2, VP3. Note that, as from this part N means standard normal distribution, SH means symmetric heavy tailed distribution, AL means asymmetric light distribution and AH means asymmetric heavy tailed distribution. Like Table 3.2; in Table 3.3 x has symmetric heavy tailed distribution, in Table 3.4 x has asymmetric light tailed distribution and in Table 3.5 x has asymmetric heavy tailed distribution, for each table all error distribution and variance pattern combinations are given.

Table 3.2 Estimated ratios of standard errors (R), x has standard normal distribution

x	ε	VP	TS	LMS	LTS	LTA	LAV	M	Preference
N	**N**	**1**	0.919	0.448	0.470	0.661	0.849	0.904	**OLS, TS, M**
		2	2.635	2.321	2.284	2.166	2.082	1.965	**TS, LMS, LTS**
		3	341.571	190.389	225.836	364.038	372.102	405.777	**M, LAV, LTA**
N	**SH**	**1**	3.113	2.099	2.231	3.069	3.281	3.438	**M, LAV, TS**
		2	27.818	27.056	26.813	25.316	22.547	20.470	**TS, LMS, LTS**
		3	167.779	126.655	146.239	206.302	198.729	206.640	**M, LTA, LAV**
N	**AL**	**1**	1.140	0.616	0.650	0.858	1.021	1.093	**TS, M, LAV**
		2	3.378	2.940	2.900	2.694	2.580	2.443	**TS, LMS, LTS**
		3	45.149	27.857	32.941	48.879	48.816	53.298	**M, LTA, LAV**
N	**AH**	**1**	6.351	4.542	4.913	6.373	6.463	6.810	**M, LAV, LTA**
		2	13.805	13.575	13.459	12.589	11.076	10.129	**TS, LMS, LTS**
		3	766.003	580.291	695.664	957.972	908.334	943.310	**LTA, M, LAV**

Table 3.3 Estimated ratios of standard errors (R), x has symmetric heavy tailed distribution

x	ε	VP	TS	LMS	LTS	LTA	LAV	M	Preference
SH	**N**	**1**	0.810	0.333	0.369	0.546	0.843	0.844	**OLS, M, LAV**
		2	42.546	47.530	45.358	35.509	2.802	7.688	**LMS, LTS, TS**
		3	54.441	30.938	38.235	61.726	67.366	80.365	**M, LAV, LTA**
SH	**SH**	**1**	2.305	1.389	1.543	2.157	2.392	2.699	**M, LAV, LTA**
		2	105.626	125.078	119.215	93.373	7.865	17.458	**LMS, LTS, TS**
		3	36.392	28.696	35.662	50.090	49.863	53.307	**M, LTA, LAV**
SH	**AL**	**1**	0.984	0.444	0.497	0.672	0.957	1.025	**M, OLS, TS**
		2	55.746	61.844	58.900	46.635	3.595	9.915	**LMS, LTS, TS**
		3	2683.368	1588.837	2195.021	2984.511	3242.144	3855.581	**M, LAV, LTA**
SH	**AH**	**1**	9.644	5.986	6.652	9.054	10.133	11.143	**M, LAV, TS**
		2	286.907	345.789	331.052	253.054	20.789	32.692	**LMS, LTS, TS**
		3	70.066	57.076	74.967	95.056	93.989	101.052	**M, LTA, LAV**

Table 3.4 Estimated ratios of standard errors (R), x has asymmetric light tailed distribution

x	ε	VP	TS	LMS	LTS	LTA	LAV	M	Preference
AL	**N**	**1**	0.882	0.412	0.435	0.609	0.836	0.888	**OLS, M, TS**
		2	7.094	6.495	6.400	6.017	4.110	3.890	**TS, LMS, LTS**
		3	52.088	26.584	32.247	54.461	57.289	65.142	**M, LAV, LTA**
AL	**SH**	**1**	5.039	3.219	3.481	4.803	5.323	5.692	**M, LAV, TS**
		2	22.366	23.513	23.310	21.087	11.940	11.359	**LMS, LTS, TS**
		3	70.267	52.467	62.440	85.326	85.467	93.702	**M, LAV, LTA**
AL	**AL**	**1**	1.073	0.554	0.596	0.735	0.974	1.064	**TS, M, OLS**
		2	8.903	7.928	7.859	7.522	5.176	4.378	**TS, LMS, LTS**
		3	131.339	78.411	87.787	135.082	140.254	158.505	**M, LAV, LTA**
AL	**AH**	**1**	7.651	5.356	5.867	7.294	7.799	8.408	**M, LAV, TS**
		2	42.115	43.384	43.370	39.426	23.143	21.073	**LMS, LTS, TS**
		3	221.962	169.376	210.072	263.158	261.574	288.587	**M, LTA, LAV**

Table 3.5 Estimated ratios of standard errors (R), x has asymmetric heavy tailed distribution

x	ε	VP	TS	LMS	LTS	LTA	LAV	M	Preference
AH	**N**	**1**	0.789	0.316	0.349	0.518	0.836	0.879	**OLS, M, LAV**
		2	101.854	118.008	112.109	86.783	4.365	13.624	**LMS, LTS, TS**
		3	62.716	33.411	39.267	64.886	75.581	98.705	**M, LAV, LTA**
AH	**SH**	**1**	2.500	1.465	1.668	2.349	2.783	3.036	**M, LAV, TS**
		2	371.416	464.168	444.411	330.557	17.186	2.039	**LMS, LTS, TS**
		3	256.397	202.557	260.999	329.487	336.789	382.507	**M, LAV, LTA**
AH	**AL**	**1**	0.961	0.431	0.480	0.623	0.985	1.061	**M, OLS, LAV**
		2	169.337	193.832	189.471	143.206	7.907	21.629	**LMS, LTS TS**
		3	59.732	34.452	44.542	62.438	66.319	86.811	**M, LAV, LTA**
AH	**AH**	**1**	4.661	2.760	3.102	4.243	4.843	5.481	**M, LAV, TS**
		2	502.030	622.423	587.302	443.276	22.858	38.356	**LMS, LTS, TS**
		3	44.771	35.256	47.114	55.663	57.989	67.626	**M, LAV, LTA**

In terms of relative efficiency, robust regression estimators generally performs better. In Table 3.2, when both x and are ε normal with homoscedasticity, OLS has smaller standard error. Additionally in this table, that is when x has normal distribution, if the model is homoscedastic, M, TS and LAV generally gives more successful results. Furthermore, with VP2, TS, LMS and LTS; with VP3, LTA, M and LAV perform better.

In Table 3.3, where x has symmetric heavy tailed distribution, again there are some exceptions that OLS has smaller standard errors. For the rest, when the model is homoscedastic, M, LAV or TS get better results. Also it is seen from the table; with VP2, LMS, LTS and TS have smaller standard errors and with VP3, M, LAV and LTA perform better.

In Table 3.4, especially when ε has normal distribution OLS has smaller standard errors. In general, with VP1, M, LAV or TS are appropriate estimators. Furthermore, with VP2, LMS, LTS and TS; with VP3, M, LAV or LTA can be chosen to get smaller standard errors.

In Table 3.5, where x has asymmetric heavy tailed distribution, when there is homoscedasticity, the relative efficiency of M, LAV or TS estimators are better. With VP2, LMS, LTS and TS; with VP3, M, LAV and LTA are the estimators that have better relative efficiency.

3.2.2 Actual Significance Levels Results

In terms of actual significance levels, the results are given in Table 3.6, Table 3.7, Table 3.8 and Table 3.9. The organization of the tables is similar with the relative efficiency results. For each case, the actual significance levels that are falling in the stated interval (0.025, 0.075) are marked. The three best robust regression estimators that corresponds to marked actual significance levels are given as preference in last columns of the tables. The most important criteria is falling in the interval (0.025, 0.075). So, in some cases, there are two marked actual significance levels and two preferences, because there is no third actual significance level between 0.025 and 0.075. Similarly sometimes there is only one preference. Moreover, there are cases that has no preference since none of the actual significance levels falls in the stated interval. Note that Bradley's conservative criteria requires to stay within the bounds 0.045 and 0.055, when testing at the 0.05 level. But, in this simulation study, Bradley's liberal criteria is considered, so the actual significance levels should fall in the interval (0.025, 0.075) for robustness (Bradley, 1978).

Furthermore, Figure 3.5, Figure 3.6, Figure 3.7 and Figure 3.8 shows how the actual significance levels of more appropriate robust regression estimators spread around the nominal 0.05 level.

Table 3.6 Actual significance levels, x has normal distribution

x	ε	VP	TS	LMS	LTS	LTA	M	LAV	OLS	Preference
N	N	1	0.0440	0.0008	0.0008	0.0180	0.0436	0.0348	0.0492	OLS,TS,M
		2	0.0664	0.0058	0.0092	0.0394	0.0612	0.0626	0.3612	LTA, M, LAV
		3	0.0310	0	0.0008	0.0114	0.0300	0.0210	0.0098	TS, M
N	SH	1	0.0436	0.0010	0.0018	0.0190	0.0434	0.0354	0.0516	OLS, TS, M
		2	0.0608	0.0036	0.0070	0.0342	0.0554	0.0586	0.3026	M, LAV,TS
		3	0.0344	0	0.0018	0.0148	0.0334	0.0242	0.0138	TS, M
N	AL	1	0.0436	0.0004	0.0012	0.0212	0.0436	0.0342	0.0474	OLS, TS, M
		2	0.0652	0.0068	0.0086	0.0416	0.0622	0.0612	0.3442	LTA,LAV,M
		3	0.0280	0.0002	0.0004	0.0140	0.0312	0.0184	0.0114	M, TS
N	AH	1	0.0400	0.0010	0.0018	0.0188	0.0412	0.0308	0.0478	OLS, M, TS
		2	0.0606	0.0046	0.0074	0.0400	0.0580	0.0594	0.2800	M,LAV,LTA
		3	0.0338	0.0004	0.0012	0.0148	0.0336	0.0240	0.0200	TS, M

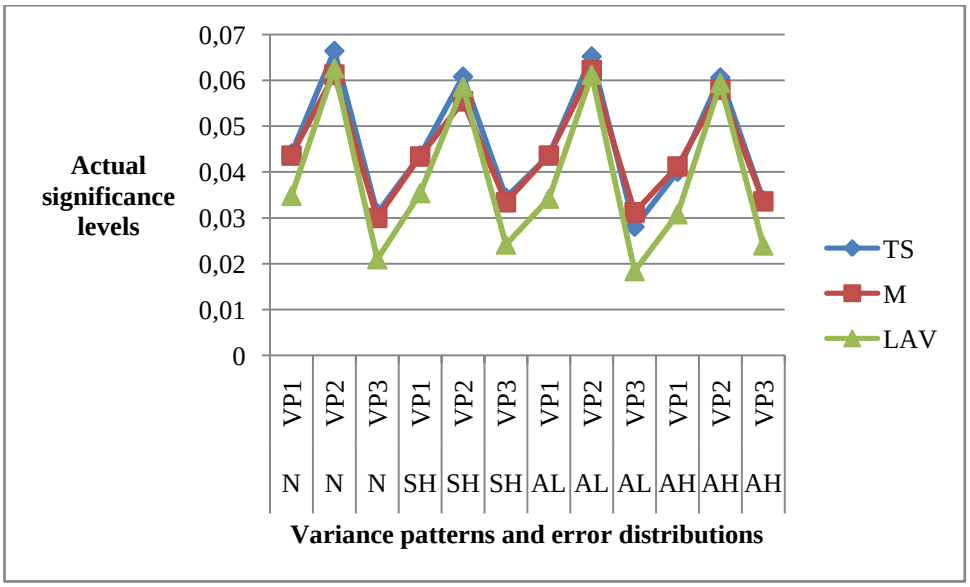


Figure 3.5 Actual significance levels of the three best robust regression estimators when x has normal distribution

Table 3.7 Actual significance levels, x has symmetric heavy tailed distribution

x	ε	VP	TS	LMS	LTS	LTA	M	LAV	OLS	Preference
SH	**N**	**1**	0.0428	0.0004	0.0012	0.0154	0.0440	0.0316	0.0462	**OLS, M, TS**
		2	0.0714	0.0068	0.0098	0.0434	0.0764	0.0862	0.6880	**LTA, TS**
		3	0.0210	0	0.0002	0.0068	0.0178	0.0098	0.0014	-
SH	**SH**	**1**	0.0398	0.0008	0.0012	0.0156	0.0428	0.0314	0.0650	**M, TS, OLS**
		2	0.0626	0.0046	0.0068	0.0318	0.0722	0.0808	0.6454	**TS, LTA, M**
		3	0.0274	0.0002	0.0006	0.0084	0.0196	0.0110	0.0042	**TS**
SH	**AL**	**1**	0.0468	0.0006	0.0018	0.0194	0.0436	0.0328	0.0518	**OLS, TS, M**
		2	0.0694	0.0056	0.0078	0.0364	0.0720	0.0812	0.6958	**LTA, TS**
		3	0.0220	0.0004	0.0004	0.0046	0.0118	0.0078	0.0014	-
SH	**AH**	**1**	0.0454	0.0006	0.0026	0.0184	0.0440	0.0332	0.0620	**TS, M, OLS**
		2	0.0672	0.0054	0.0076	0.0360	0.0760	0.0840	0.6332	**LTA, TS**
		3	0.0338	0.0006	0.0010	0.0058	0.0216	0.0122	0.0070	**TS**

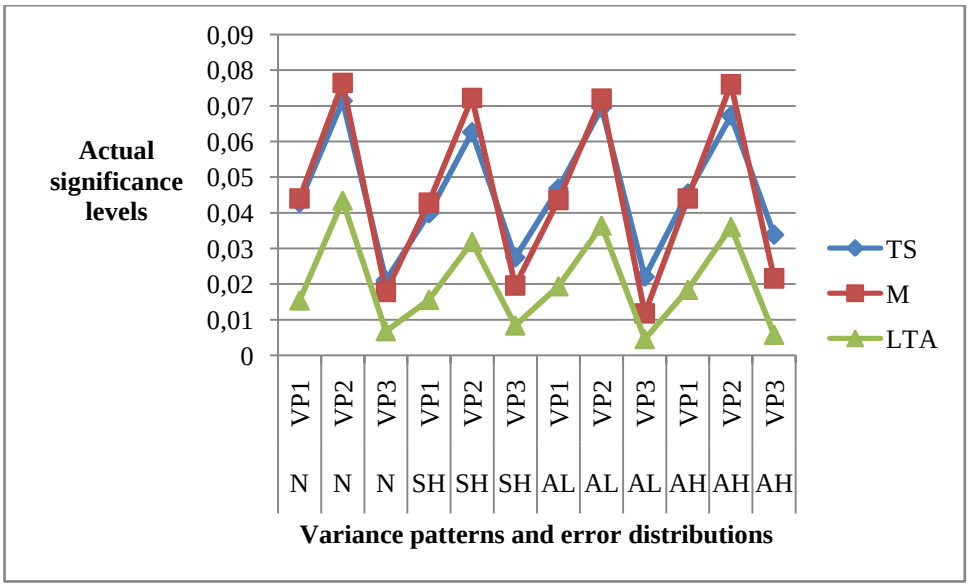


Figure 3.6 Actual significance levels of the three best robust regression estimators when x has symmetric heavy tailed distribution

Table 3.8 Actual significance levels, x has asymmetric light tailed distribution

x	ε	VP	TS	LMS	LTS	LTA	M	LAV	OLS	Preference
AL	**N**	**1**	0.0424	0.0002	0.0012	0.0152	0.0414	0.0300	0.0470	**OLS, TS, M**
		2	0.0610	0.0038	0.0070	0.0402	0.0648	0.0668	0.5330	**LTA, TS, M**
		3	0.0300	0	0.0002	0.0110	0.0322	0.0184	0.0060	**M, TS**
AL	**SH**	**1**	0.0450	0.0014	0.0020	0.0192	0.0452	0.0316	0.0496	**OLS, M, TS**
		2	0.0592	0.0042	0.0086	0.0372	0.0642	0.0632	0.4680	**TS, LTA, LAV**
		3	0.0308	0.0006	0.0010	0.0120	0.0296	0.0186	0.0160	**TS, M**
AL	**AL**	**1**	0.0470	0.0004	0.0010	0.0204	0.0486	0.0308	0.0484	**M, OLS, TS**
		2	0.0610	0.0040	0.0068	0.0384	0.0638	0.0658	0.5174	**TS, LTA, M**
		3	0.0302	0.0004	0.0012	0.0144	0.0320	0.0186	0.0072	**M, TS**
AL	**AH**	**1**	0.0470	0.0002	0.0018	0.0194	0.0458	0.0306	0.0536	**TS, OLS, M**
		2	0.0596	0.0038	0.0066	0.0404	0.0628	0.0642	0.4482	**TS, LTA, M**
		3	0.0372	0.0002	0.0022	0.0142	0.0334	0.0220	0.0168	**TS, M**

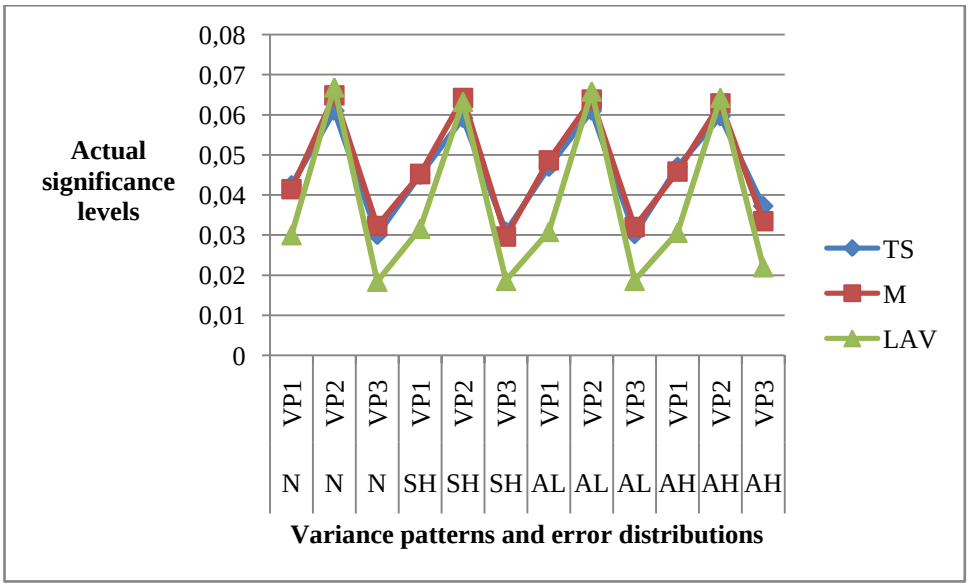


Figure 3.7 Actual significance levels of the three best robust regression estimators when x has asymmetric light tailed distribution

Table 3.9 Actual significance levels, x has asymmetric heavy tailed distribution

x	ε	VP	TS	LMS	LTS	LTA	M	LAV	OLS	Preference
AH	**N**	**1**	0.0466	0.0002	0.0022	0.0166	0.0480	0.0352	0.0478	**M, OLS, TS**
		2	0.0720	0.0064	0.0086	0.0414	0.0770	0.0842	0.7274	**LTA, TS**
		3	0.0238	0	0	0.0070	0.0184	0.0086	0.0008	-
AH	**SH**	**1**	0.0472	0.0004	0.0020	0.0182	0.0476	0.0316	0.0672	**M, TS, OLS**
		2	0.0658	0.0050	0.0078	0.0368	0.0718	0.0810	0.6918	**LTA, TS, M**
		3	0.0288	0.0002	0.0012	0.0082	0.0220	0.0124	0.0056	**TS**
AH	**AL**	**1**	0.0402	0.0004	0.0012	0.0164	0.0434	0.0320	0.0532	**OLS, M, TS**
		2	0.0630	0.0060	0.0078	0.0358	0.0698	0.0816	0.7188	**TS, LTA, M**
		3	0.0220	0.0002	0.0006	0.0064	0.0168	0.0090	0.0012	-
AH	**AH**	**1**	0.0454	0.0012	0.0022	0.0186	0.0436	0.0360	0.0648	**TS, M, LAV**
		2	0.0620	0.0052	0.0072	0.0350	0.0649	0.0800	0.6762	**TS, M, LTA**
		3	0.0304	0.0002	0.0006	0.0100	0.0222	0.0124	0.0074	**TS**

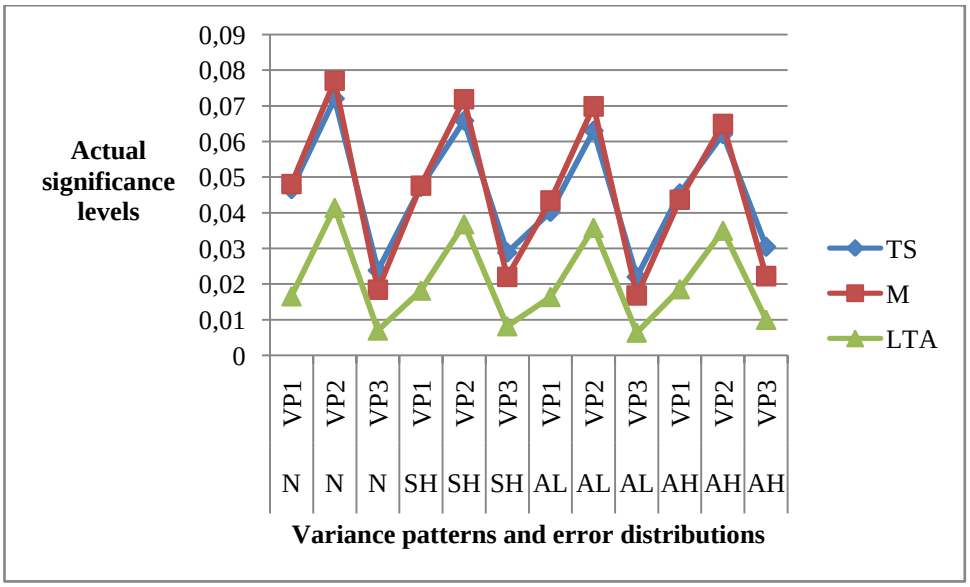


Figure 3.8 Actual significance levels of the three best robust regression estimators when x has asymmetric heavy tailed distribution

Robust regression estimators are more successful in having actual significance levels in the interval (0.025, 0.075) when they are compared with OLS. In Table 3.6, where x has normal distribution, TS and M have their actual significance levels in the stated interval in all cases. Also OLS has actual significance levels that are close to

the nominal level when there is homoscedasticity. With VP1 and VP2, LAV gives results within the bounds 0.025 and 0.075. Finally, LTA can keep its results within the interval with VP2.

In Table 3.7, where x has symmetric heavy tailed distribution, TS's actual significance levels fall in the interval in almost all cases. M, LAV or OLS can be preferred with homoscedasticity and LTA performs better with VP2.

In Table 3.8, x has asymmetric but light tailed distribution. It is seen that, TS and M estimators keep actual significance levels in the interval (0.025, 0.075) in all cases. OLS can only be used when there is homoscedasticity. LAV performs well with VP1 and VP2, besides LTA performs well with VP2.

In Table 3.9, x has asymmetric and heavy tailed distribution. TS estimator has actual significance levels that falling in the interval in almost all cases, and M estimator is the second estimator that performs well. LAV and OLS can be preferred with VP1, LTA can be preferred with VP2. Also it is easily seen from the table, with VP2, although ε has normal distribution, the actual significance level of OLS is approximately 0.73 when testing at 0.05 level.

3.3 Real Data Example

In this part, the robust regression techniques and OLS are applied to four different real data sets. The real data sets are obtained from http://www.freewebs.com/tedstats. Some properties of the real data is given in Table 3.10 and their histograms are given in Figure 3.9, Figure 3.10, Figure 3.11 and Figure 3.12. The sample sizes, regression models, number of replications and comparison criteria are the same as simulation study in previous part.

Table 3.10 The descriptive statistics of real data sets

Data Name	N	mean	median	0.2 trimmed mean	standard deviation	skewness	kurtosis
Carnegie Data							
Total AA Degrees	**2859**	**268.98**	**114**	**140.83**	**451.96**	**4.98**	**47.31**
Total MA Degrees	**1845**	**353.51**	**125**	**163.58**	**611.17**	**4.01**	**27.03**
Total BA Degrees	**2417**	**664.23**	**254**	**304.11**	**1164.52**	**3.63**	**19.88**
Standardized Test							
GRE Total	**4637**	**1055.44**	**1060**	**1062.34**	**185.104**	**-0.25**	**2.84**

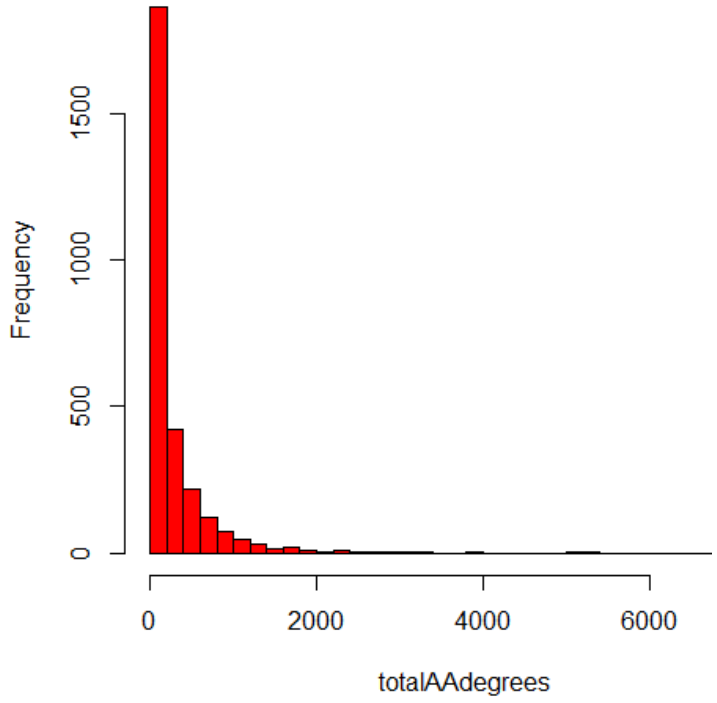


Figure 3.9 Histogram of total AA degrees

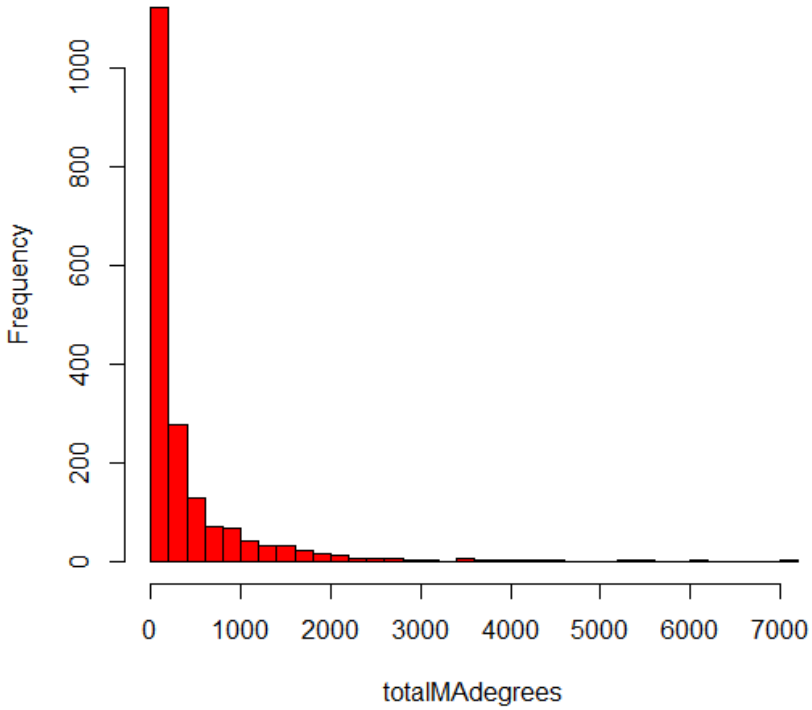


Figure 3.10 Histogram of total MA degrees

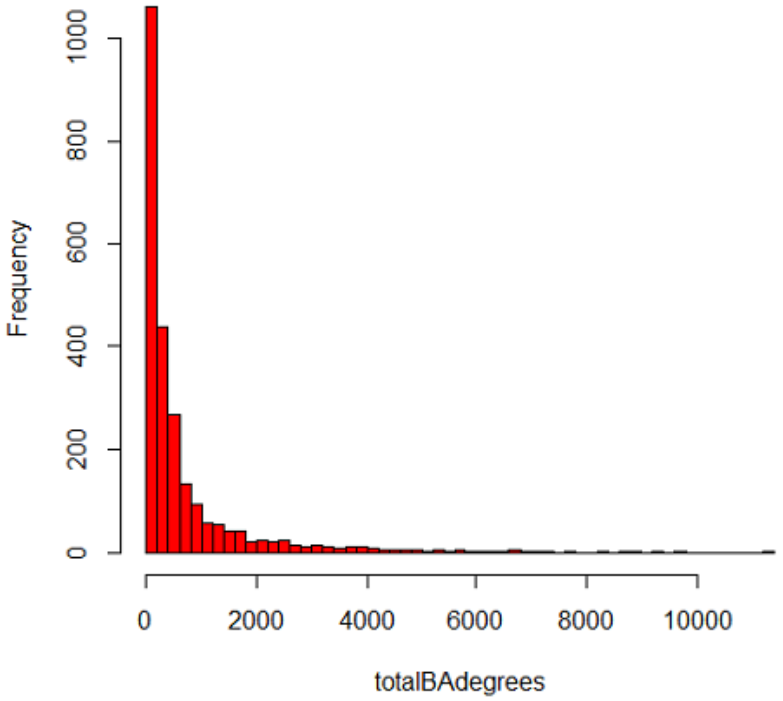


Figure 3.11 Histogram of total BA degrees

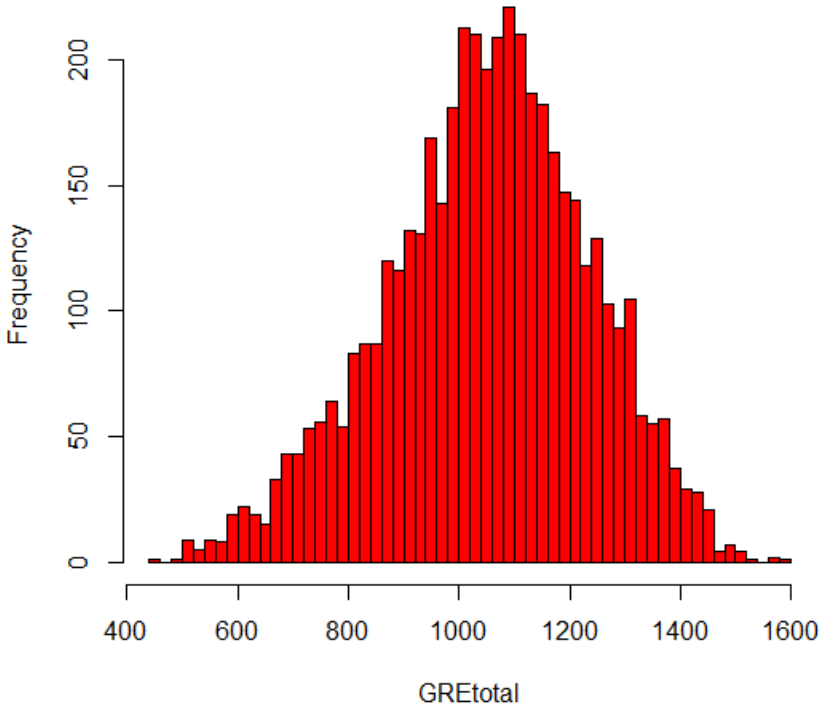


Figure 3.12 Histogram of GRE total

The results of the relative efficiency and actual significance levels comparisons are given in Table 3.11 and Table 3.12 respectively.

Table 3.11 Estimated ratios of standard errors (R) with real data

	TS	**LMS**	**LTS**	**LTA**	**LAV**	**M**	**Preference**
Total AA Degrees	1.866	1.391	1.439	1.402	1.559	1.705	TS, M, LAV
Total MA Degrees	1.865	1.352	1.409	1.352	1.475	1.612	TS, M, LAV
Total BA Degrees	1.984	1.413	1.503	1.554	1.674	1.755	TS, M, LAV
GRE Total	0.921	0.448	0.469	0.901	0.912	0.905	OLS, TS, LAV

Table 3. 12 Actual significance levels with real data

	TS	LMS	LTS	LTA	LAV	M	OLS	Preference
Total AA Degrees	0.0396	0.0010	0.0030	0.0280	0.0376	0.0406	0.0538	OLS, M, TS
Total MA Degrees	0.0494	0.0008	0.0024	0.0280	0.0418	0.0464	0.0604	TS, M, OLS
Total BA Degrees	0.0432	0.0014	0.0054	0.0258	0.0426	0.0456	0.0582	M, TS, LAV
GRE Total	0.0416	0.0008	0.0036	0.0260	0.0384	0.0456	0.0572	M, OLS, TS

In Table 3.11, the greatest three R ratios are marked. The greatest values are obtained with the same estimators in all real data sets. The robust regression estimators that perform better are TS, LAV and M. With AA Degrees, MA Degrees and BA degrees R ratios are greater than one, that is robust regression estimator has smaller standard error. This is an expected result because AA Degrees, MA Degrees and BA Degrees are asymmetric data sets. On the other side, since GRE Total is symmetric, the R ratios are smaller than one, which means OLS performs better with this data set.

In Table 3.12, three actual significance levels that are closest to the nominal level are marked. These actual significance levels are generally obtained with TS, M and OLS. With all data sets, TS, LTA, LAV, M and OLS estimator have actual significance levels that are falling in the interval (0.025, 0.075) which is concerned with Bradley's liberal criteria. Although OLS stays within the liberal bounds, its actual significance levels do not fall in the interval (0.045, 0.055) which is formed according to Bradley's conservative interval.

CHAPTER FOUR
CONCLUSION

In this thesis, problems in regression analysis that are caused by violation of normality and homoscedasticity assumptions are investigated. At first, this problem is stated. Then new approaches which are called robust are defined. With a simulation study their performance is compared with respect to relative efficiency over OLS and actual significance levels.

In robust regression techniques, as in robust theory, it is desired to reduce the effects of outliers. For example, when location parameters are considered, median or trimmed mean are more robust against outliers when they are compared with arithmetic mean. With a similar reasoning, instead of squared residuals, by trimming or by developing some computations that use median, it is possible to get robust estimators in regression analysis.

In the first simulation study, in order to compare robust regression estimators in terms of relative efficiency, estimators that have R ratio greater than one are preferred. On the other hand, in order to compare robust regression estimators in terms of actual significance levels, estimators that have their actual significance level in the interval (0.025, 0.075) are preferred.

At the beginning, it is important to note that there is no estimator that can be best for all conditions. The appropriate robust regression estimator should be preferred according to distribution of x and ε, plus the level of heteroscedasticity.

In terms of relative efficiency, robust regression estimators generally compete well with the OLS estimator. The basic exceptions are the cases that the error term has normal distribution with VP1. It means when the error term ε has normal distribution with homoscedasticity, whatever the distribution of x, the standard error of OLS is smaller. However, it is easily seen that, when x has symmetric heavy tailed and ε has asymmetric light tailed distribution with VP3, the OLS estimator

can have a standard error which is even 3855 times of some robust regression estimator.

If continued with relative efficiency, when there is homoscedasticity, the best choice is M-regression. However, in some cases where the error term has asymmetric and light tailed distribution TS estimator is more appropriate.

With VP2, where the error term ε has larger variance than the extreme x values, TS and LMS estimators can be chosen. When x has symmetric or asymmetric heavy tailed distribution, whatever ε has, LMS estimator has the smallest standard error. Also LTS estimator can be a choice since its results are very close to the results of LMS estimator. Furthermore, when x has normal or asymmetric light tailed distribution, TS can be chosen.

Finally, with VP3, where the error term ε has smaller variance than extreme x values, the most appropriate estimator is M-regression in almost all distribution cases.

In short, 48 conditions which are formed by combining different distributions and variance patterns are investigated. The greatest R ratio is obtained by M-regression 24 times, with LMS estimator 11 times with a close follower either TS or M-regression, with TS 8 times. Also as it is seen, in four condition the most appropriate estimator is OLS and in one condition it is LTA.

In terms of actual significance levels ($\hat{\alpha}$), when there is homoscedasticity, results of OLS fall in the interval (0.025, 0.075). Also, whatever the distributions of x and ε are; TS, M and LAV are the robust regression estimators that have $\hat{\alpha}$ values between 0.025 and 0.075. Neither of them is best in all conditions, however with VP1, TS or M can be appropriate.

With VP2 and when x has normal distribution the best choice can be one of M or LAV estimators. Furthermore, if x has asymmetric heavy tailed, symmetric heavy

tailed or asymmetric light tailed distribution, the preferences are TS and LTA estimators.

With the last heteroscedasticity condition VP3, in four cases no preference is given. When x has heavy tailed distribution and ε has normal or asymmetric heavy tailed distribution none of the robust regression methods used in this study has their results in the interval (0.025, 0.075). Except these cases, when x has heavy tailed distribution TS can be preferred. Additionally if x has normal or asymmetric light tailed distribution TS or M can be chosen in terms of actual significance levels.

In 48 conditions that are investigated, TS's results fall in the stated interval 44 times and M follows this score 33 times. Furthermore LAV stays within the bounds 0.025 and 0.075 24 times, LTA does 16 times.

Results of the real data example support the results of simulation study. With asymmetric data sets TS, M or LAV estimators can be chosen and with the symmetric data set M or TS can be preference, besides it is possible to use OLS.

Although LMS and LTS have larger R values in some cases, their actual significance levels never fell in the interval (0.025, 0.075). Hence, instead of using LMS or LTS, using an estimator that has larger R values, plus has its actual significance levels close to the nominal level is more appropriate. Through the robust regression estimators that are used in this thesis, LMS and LTS are not preferred.

To summarize, deciding the appropriate robust regression estimator is concerned with both the level of heteroscedasticity and distributions of x and ε. If the assumptions violated M-regression or Theil-Sen estimator can be appropriate instead of conventional technique. Although M and TS seem to be good choices generally, there is no technique that performs best in all conditions.

REFERENCES

Birkes, D., & Dodge, Y. (1993). *Alternative methods of regression.* New York: Wiley.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144-152.

Box, G. E. P. (1953). Non- normality and tests on variances. *Biometrika, 40,* 318-335.

Croux, C., Rousseeuw, P. J., & Hössjer, O. (1994). Generalized S-estimators. *Journal of The American Statistical Association, 89,* 1271- 1281.

David, H. A. (1995). First (?) occurrence of common terms in mathematical statistics. *The American Statisticians, 49,* 121-133.

Dietz, E. J. (1989). Teaching regression in a nonparametric statistics course. *The American Statisticians, 43,* 35-40.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of The American Statistical Association, 69,* 383- 393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. , & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley.

Hawkins, D. M., & Olive, D. (1999). Applications and algorithms for least trimmed sum of absolute deviations regression. *Computational Statistics & Data Analysis, 32,* 119-134.

Hawkins, D. M., & Olive, D. (2002). Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm. *Journal of The American Statistical Association, 97,* 136-147.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1985). *Exploring data tables, trends and shapes.* New York: Wiley.

Micceri, T. (2014). *Real data*. Retrieved June 24, 2014, from http://www.freewebs.com/tedstats.

Rousseeuw, P. J. (1984). Least median of squares. *Journal of The American Statistical Association, 79,* 871-880.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression & outlier detection*. New York: Wiley.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of The American Statistical Association, 63,* 1379- 1389.

Stromberg, A. J., Hössjer, O., & Hawkins, D. M. (2000). The least trimmed differences regression estimator and alternatives. *Journal of The American Statistical Association, 95,* 853-874.

Theil, H. (1950). A rank-invariant method of linear an polynomial regression analysis. *Indagationes Mathematicae, 12,* 85-91.

Wilcox, R. R. (1996a). Estimation in the simple linear regression model when there is heteroscedasticity of unknown form. *Communications in Statistics- Theory and Methods, 25,* 1305-1324.

Wilcox, R. R. (1996b). Confidence intervals for the slope of a regression line when the error term has nonconstant variance. *Computational Statistics& Data Analysis, 22,* 89-98.

Wilcox, R. R. (1996c). A review of some recent developments in robust regression. *British Journal of Mathematical and Statistical Psychology, 49,* 253-274.

Wilcox, R. R. (1998a). The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology, 51,* 1-39.

Wilcox, R. R. (1998b). Simulation results on extensions of the Theil-Sen estimator. *Communications in Statistics- Theory and Methods, 27,* 1117-1126.

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods, substantially improving power and accuracy*(2nd ed.).New York: Springer.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Amsterdam: Elsevier/ Academic Press.

Wilcox, R. R., & Keselman, H. J. (2012). Modern regression methods that can substantially increase power and provide a more accurate understanding of associations. *European Journal Personality, 26,* 165-174.